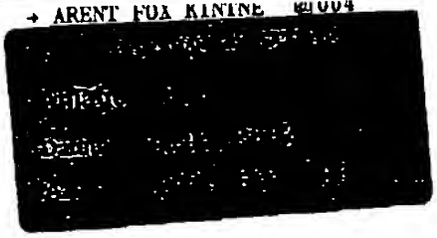

ATLAS of PROTEIN SEQUENCE and STRUCTURE

Volume 5

SUPPLEMENT 3
1978

Margaret O. Dayhoff

National Biomedical Research Foundation
Georgetown University Medical Center
3900 Reservoir Road, N.W.
Washington, D.C. 20007
Tel. 202–625-2121

22 A Model of Evolutionary Change in Proteins

M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt

In the eight years since we last examined the amino acid exchanges seen in closely related proteins,[1] the information has doubled in quantity and comes from a much wider variety of protein types. The matrices derived from these data that describe the amino acid replacement probabilities between two sequences at various evolutionary distances are more accurate and the scoring matrix that is derived is more sensitive in detecting distant relationships than the one that we previously derived.[2,3] The method used in this chapter is essentially the same as that described in the *Atlas*, Volume 3[4] and Volume 5.[1]

Accepted Point Mutations

An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of a mutation in the portion of the gene template producing one amino acid of a protein; the second is the acceptance of the mutation by the species as the new predominant form. To be accepted, the new amino acid usually must function in a way similar to the old one: chemical and physical similarities are found between the amino acids that are observed to interchange frequently.

Any complete discussion of the observed behavior of amino acids in the evolutionary process must consider the frequency of change of each amino acid to each other one and the propensity of each to remain unchanged. There are 20 X 20 = 400 possible comparisons. To collect a useful amount of information on these, a great many observations are necessary. The body of data used in this study includes 1,572 changes in 71 groups of closely related proteins appearing in the *Atlas* volumes through Supplement 2.

The mutation data were accumulated from the phylogenetic trees and from a few pairs of related sequences. The sequences of all the nodal common ancestors in each tree are routinely generated. Consider, for example, the much simplified artificial phylogenic tree of Figure 78. The matrix of accepted point mutations calculated from this tree is shown in Figure 79. We have assumed that the likelihood of amino acid X replacing Y is the same as that of Y replacing X, and hence 1 is entered in box YX as well as in box XY. This assumption is reasonable, because this likelihood should depend on the product of the frequencies of occurrence of the two amino acids and on their chemical and physical similarity. As a consequence of this assumption, no change in amino acid frequencies over evolutionary distance will be detected.

By comparing observed sequences with inferred ancestral sequences, rather than with each other, a sharper

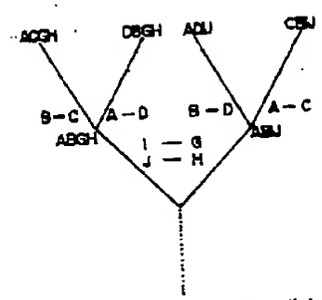

Figure 78. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

Figure 79. Matrix of accepted point mutations derived from the tree of Figure 78.

picture of the acceptable point mutations is obtained. In the first amino acid position of the illustration in Figure 78, A changes to D and A changes to C, but C and D do not directly interchange. If we had compared the observed sequences directly, we would have inferred the change of C to D also. (In practice, some of the positions in the nodal sequences are blank [ambiguous]. For these, we have treated the changes statistically, distributing them among all observed alternatives.)

The total numbers of accepted point mutations observed between closely related sequences from 34 superfamilies, grouped into 71 evolutionary trees, are shown in Figure 80. In order to minimize the occurrence of changes caused by successive accepted mutations at one site, the sequences within a tree were less than 15% different from one another and ancestral sequences were even closer. Of the 190 possible exchanges shown in Figure 80, 35 were never observed. These usually involved the amino acids that occur infrequently and are not highly mutable and exchanges where more than one nucleotide of the codon must change. Of the 1,572 exchanges the largest number, 83, was observed between Asp and Glu, two chemically very similar amino acids with codons differing by one nucleotide. About 20% of the interchanges, far more than one would expect for such similar sequences, involved amino acids whose codons differed by more than one nucleotide. Presumably, in any one tree, changes at some of the amino acid positions are rejected by selection and multiple changes at the mutable sites are favored. Many of the changes expected from the mutations of one nucleotide in a codon are seldom observed. Presumably these mutations have occurred often but have been rejected by natural selection acting on the proteins. For example, there were no exchanges between Gly and Trp.

Mutability of Amino Acids

A complete picture of the mutational process must include a consideration of the amino acids that did not

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
A	Ala																			
R	Arg	30																		
N	Asn	109	17																	
D	Asp	154	0	532																
C	Cys	33	10	0	0															
Q	Gln	93	120	50	76	0														
E	Glu	266	0	94	831	0	422													
G	Gly	579	10	156	162	10	30	112												
H	His	21	103	226	43	10	243	23	10											
I	Ile	66	30	36	13	17	8	35	0	3										
L	Leu	95	17	37	0	0	75	15	17	40	253									
K	Lys	57	477	322	65	0	147	104	60	23	43	39								
M	Met	25	17	0	0	0	20	7	7	0	57	207	90							
F	Phe	20	7	7	0	0	0	0	17	20	90	167	0	17						
P	Pro	345	87	27	10	10	93	40	49	50	7	43	43	4	7					
S	Ser	772	137	432	98	117	47	85	450	26	20	32	168	20	40	269				
T	Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696			
W	Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0		
Y	Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6	
V	Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17

Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val

Figure 80. Numbers of accepted point mutations (X10) accumulated from closely related sequences. Fifteen hundred and seventy-two exchanges are shown. Fractional exchanges result when ancestral sequences are ambiguous.

change, as well as those that did. For this we need to know the probability that each amino acid will change in a given small evolutionary interval. We call this number the "relative mutability" of the amino acid.

In order to compute the relative mutabilities of the amino acids, we simply count the number of times that each amino acid has changed in an interval and the number of times that it has occurred in the sequences and thus has been subject to mutation. The relative mutability of each amino acid is proportional to the ratio of changes to occurrences. Figure 81 illustrates the computation for a simple case in which B changes relatively often, A less often, and D never.

Aligned sequences	A D A A D B		
	A	B	D
Amino acids	1	1	0
Changes	3	1	2
Frequency of occurrence (total composition)	.33	1	0
Relative mutability			

Figure 81. Sample computation of relative mutability. The two aligned sequences may be two experimentally observed sequences or an observed sequence and its inferred ancestor.

In calculating relative mutabilities from many trees, the information from sequences of different lengths and evolutionary distances is combined. Each relative mutability is still a ratio. The numerator is the total number of changes of this amino acid on all branches of all protein trees considered. The denominator is the total exposure of the amino acid to mutation, that is, the sum for all branches of its local frequency of occurrence multiplied by the total number of mutations per 100 links for that branch.

The relative mutabilities of the amino acids are shown in Table 21. On the average, Asn, Ser, Asp, and Glu are most mutable and Trp and Cys are least mutable.

The immutability of cysteine is understandable. Cysteine is known to have several unique, indispensible functions. It is the attachment site of heme groups in cytochrome and of FeS clusters in ferredoxin. It forms cross-links in other proteins such as chymotrypsin or ribonuclease. It seldom occurs without having an important function.

The substitution of one of the larger amino acids of distinctive shape and chemistry for any other is rather uncommon. At the other extreme, the low mutability of glycine must be due to its unique smallness that is advantageous in many places. Even though serine sometimes functions in the active center, it much more often performs a function of lesser importance, easily mimicked by several other amino acids of similar physical and chemical properties. On the average it is highly mutable.

Table 21
Relative Mutabilities of the Amino Acids[a]

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

[a]The value for Ala has been arbitrarily set at 100.

Amino Acid Frequencies in the Mutation Data

The relative frequencies of exposure to mutation of the amino acids are shown in Table 22. These frequencies, f_i, are approximately proportional to the average composition of each group multiplied by the number of mutations in the tree. The sum of the frequencies is 1.

Mutation Probability Matrix for the Evolutionary Distance of One PAM

We can combine information about the individual kinds of mutations and about the relative mutability of the amino acids into one distance-dependent "mutation probability matrix" (see Figure 82). An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 PAM.

Table 22
Normalized Frequencies of the Amino Acids in the Accepted Point Mutation Data

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

The nondiagonal elements have the values:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}}$$

where

A_{ij} is an element of the accepted point mutation matrix of Figure 80,
λ is a proportionality constant, and
m_j is the mutability of the jth amino acid, Table 21.

The diagonal elements have the values:

$$M_{jj} = 1 - \lambda m_j$$

Consider a typical column, that for alanine. The total probability, the sum of all the elements, must be 1. The probability of observing a change in a site containing alanine (the sum of all the elements except M_{AA}) is proportional to the mutability of alanine. The same proportionality constant, λ, holds for all columns. The individual nondiagonal terms within each column bear the same ratio to each other as do the observed mutations in the matrix of Figure 80.

The quantity $100 \times \Sigma f_i M_{ij}$ gives the number of amino acids that will remain unchanged when a protein 100 links long, of average composition, is exposed to the evolutionary change represented by this matrix. This apparent evolutionary change depends upon the choice of λ, in this case chosen so that this change is 1 mutation. Since there are almost no superimposed changes, this also represents 1 PAM of change. If λ had been four times as large, the initial matrix would have represented 4 PAMs; the discussion which follows would not be changed noticeably.

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	6	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

Simulation of the Mutational Process

For evaluating statistical methods of detecting relationships, for developing methods of measuring evolutionary distances between proteins, and for determining the accuracy of programs to construct evolutionary trees, we need to have examples of proteins at known evolutionary distances. The mutation probability matrix provides the information with which to simulate any amount of evolutionary change in an unlimited number of proteins. Further, we can start with one protein and simulate its separate evolution in duplicated genes or in divergent organisms. By considering many groups of sequences related by the same evolutionary history, a measure is readily obtained of the expected deviations due to random fluctuations in the evolutionary process.

If we only require that, on the average, one mutation takes place in the evolutionary interval of 1 PAM, we can use a simulation requiring one random number for each amino acid in the sequence, as follows: To determine the fate of the first amino acid, say Ala, a uniformly distributed random number between 0 and 1 is obtained. The first column of the mutation probability matrix (Figure 82) gives the relative probability of each possible event that may befall Ala (neglecting deletion for simplicity). If the random number falls between 0 and .9867, Ala is left unchanged. If the number is between .9867 and .9868, it is replaced with Arg, if it is between .9868 and .9872, it is replaced with Asp, and so forth. Similarly, a random number is produced for each amino acid in the sequence, and action is taken as dictated by the corresponding column of the matrix. The result is a simulated mutant sequence. Any number of these can be generated; their average distance from the original is 1 PAM although some may have no mutations and some may have two or more. The effects on the sequence of a longer period of evolution may be simulated by successive applications of the matrix to the sequence resulting from the last application.

For simulations in which a predetermined number of changes are required, a two-step process involving two random numbers for each mutation can be used. Starting with a given sequence, the first amino acid that will mutate is selected: the probability that any one will be selected is proportional to its mutability (Table 21). Then the amino acid that replaces it is chosen. The probability for each replacement is proportional to the elements in the appropriate column of Figure 82. Starting with the resultant sequence, a second mutation can be simulated, and so on, until a predetermined number of changes have been made. In this process, superimposed and back mutations may occur.

The 1 PAM matrix can be multiplied by itself N times to yield a matrix that predicts the amino acid replacements to be found after N PAMs of evolutionary change in a sequence of average composition. On the average, the results of the simulations above match the predictions of the corresponding matrices.

Mutation Probability Matrices for Other Distances

The mutation probability matrix M_1, corresponding to 1 PAM, has a number of interesting properties (see Figure 82). If, in a simulation, it is applied to a protein with the average amino acid composition given in Table 22, on the average, the composition of the resulting mutated proteins will be unchanged. Repeated applications of the matrix to proteins of any other composition will give mutants that change toward average composition; any such matrix has implicit in it some particular asymptotic composition.

There is a different mutation probability matrix for each evolutionary interval. These can be derived from the one for 1 PAM by matrix multiplication. If the 1-PAM matrix is multiplied by itself an infinite number of times, each column of the resulting matrix approaches the asymptotic amino acid composition:

$$M_\infty = \begin{matrix} f_A & f_A & f_A & f_A & \cdots \\ f_R & f_R & f_R & f_R & \cdots \\ f_N & f_N & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \end{matrix}$$

At a great distance, there is very little relationship information left in the matrix. For example, at a distance of 2,034 PAMs, all of the matrix values are within 5% of their limiting values except for the Trp-Trp element, which is 75% higher than the limit, and the Cys-Cys element, which is 11% higher.

The matrix for 0 PAMs is simply a unit diagonal; no amino acid would have changed:

$$M_0 = \begin{matrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \end{matrix}$$

The mutation probability matrix for 250 PAMs is shown in Figure 83. At this evolutionary distance, only one amino acid in five remains unchanged. However, the

amino acids vary greatly in their mutability; 55% of the tryptophans, 52% of the cysteines and 27% of the glycines would still be unchanged, but only 6% of the highly mutable asparagines would remain. Several other amino acids, particularly alanine, aspartic acid, glutamic acid, glycine, lysine, and serine are more likely to occur in place of an original asparagine than asparagine itself at this evolutionary distance! This is understandable from the data giving the preferred mutations and the relative mutabilities. Asparagine is highly mutable, therefore it changes to other amino acids. These are less mutable and may not change again. This effect is much more conspicuous in the case of methionine. Surprisingly, a methionine originally present would have changed to leucine in 20% of the cases, but would remain methionine in only 6%. Over one-third of the mutations in methionine are specifically to leucine (Figure 80). Leucine is less than one-half as mutable as methionine (Table 21).

From the series of distance-dependent mutation probability matrices, we can compute detailed answers to the question "How does the evolutionary process affect the similarity of related protein sequences?"

Estimation of Evolutionary Distance

There is a different mutation probability matrix for each evolutionary interval measured in PAMs. For each such matrix, we can calculate the percentage of amino acids that will be observed to change on the average in the interval by the formula:

$$100\left(1 - \sum_i f_i M_{ii}\right)$$

Table 23 shows the correspondance between the observed percent difference between two sequences and the evolutionary distance in PAMs. We use this scale to estimate

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	6	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala in the first sequence will contain Ala in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.

Table 23
Correspondence between Observed Differences and the Evolutionary Distance

Observed Percent Difference	Evolutionary Distance in PAMs
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

evolutionary distances from matrices of percent difference between sequences. These estimated distances were used in the computations of evolutionary trees in this book. The differences predicted for a given PAM distance differ by up to 2.5% from those that we reported in Volume 5. A more complete scale is given in Table 36 of the Appendix.

Relatedness Odds Matrix

The elements, M_{ij}, of the mutation probability matrix for each distance give the probability that amino acid j will change to i in a related sequence in that interval. The normalized frequency f_i gives the probability that i will occur in the second sequence by chance.

The terms of the relatedness odds matrix are then:

$$R_{ij} = \frac{M_{ij}}{f_i}$$

The odds matrix is symmetrical. Each term gives the probability of replacement per occurrence of i per occurrence of j.

Amino acid pairs with scores above 1 replace each other more often as alternatives in related sequences than in random sequences of the same composition whereas those with scores below 1 replace each other less often.

The information in the 250-PAM odds matrix has proven very useful in detecting distant relationships between sequences. When one protein is compared with another, position by position, one should multiply the odds for each position to calculate an odds for the whole protein. However, it is more convenient to add the logarithms of the matrix elements. The log of the 250-PAM odds matrix is shown in Figure 84.

The Chemical Meaning of Amino Acid Mutations

Patterns have been visible in the accepted point mutations since the beginning of protein sequence work. Isoleucine-valine and serine-threonine were frequently observed alternatives. It was obvious that this interchangeability had something to do with their chemical similarities. In the large amount of information that now exists, far more detailed correlations are visible, and many more functional inferences can be made.

In the log odds matrix of Figure 84, the order of the amino acids has been rearranged to show clearly the groups of chemically similar amino acids that tend to replace one another: the hydrophobic group; the aromatic group; the basic group; the acid, acid-amide group; cysteine; and the other hydrophilic residues. Some groups overlap: the basic and acid, acid-amide groups tend to replace one another to some extent, and phenylalanine interchanges with the hydrophobic group more often than chance expectation would predict. These patterns are imposed principally by natural selection and only secondarily by the constraints of the genetic code: they reflect the similarity of the functions of the amino acid residues in their weak interactions with one another in the three-dimensional conformation of proteins. Some of the properties of an amino acid residue that determine these interactions are: size, shape, and local concentrations of electric charge; the conformation of its van der Waals surface; and its ability to form salt bonds, hydrophobic bonds, and hydrogen bonds.

Computing Relationships between Sequences

We use log odds matrices as scoring matrices for detecting very distant relationships between proteins. Such scoring matrices, based ultimately on accepted point mutations, can discriminate significant relationships from

C Cys	12																				
S Ser	0	2																			
T Thr	-2	1	3																		
P Pro	-3	1	0	6																	
A Ala	-2	1	1	1	2																
G Gly	-3	1	0	-1	1	5															
N Asn	-4	1	0	-1	0	0	2														
D Asp	-5	0	0	-1	0	1	2	4													
E Glu	-5	0	0	-1	0	0	1	3	4												
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4											
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6										
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				
F Phe	-4	-3	-3	-5	-4	-5	-4	-6	-6	-6	-2	-4	-5	0	1	2	-1	9			
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
	Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp

Figure 84. Log odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of −10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and a score of +2 means that the pair would be expected to occur 1.8 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data.

random coincidences better than simpler scoring systems. Mere counts of identities and matrices based only on the changes predicted by the genetic code are not sufficiently complex. It is obvious that there is a good deal of information in the detailed nature of both the nonidentities and the identities. Certain combinations of different amino acids are positive evidence of relatedness, and others are contraindications. The log odds matrix for 250 PAMs, which we have found to be a very effective scoring matrix for detecting distant relationships, is compared with other matrices in chapter 23.

References

1. Dayhoff, M.O., Eck, R.V., and Park, C.M., *in* Atlas of Protein Sequence and Structure 1972, Vol.5, ed. Dayhoff, M.O., pp.89–99, Nat. Biomed. Res. Found., Washington, D.C., 1972
2. Schwartz, R.M., and Dayhoff, M.O., *in* Evolution of Protein Molecules, ed. Matsubara, H., and Yamanaka, T., pp.1–16, Japan Sci. Soc. Press, Tokyo, 1978
3. Schwartz, R.M., and Dayhoff, M.O., *in* Origin of Life, ed. Noda, H., pp.457–469, Center for Academic Pub. Japan/Japan Sci. Soc. Press, Tokyo, 1978
4. Dayhoff, M.O., and Eck, R.V., Atlas of Protein Sequence and Structure 1967–68, pp.33–45, Nat. Biomed. Res. Found., Silver Spring, Md., 1968

23 Matrices for Detecting Distant Relationships

R.M. Schwartz and M.O. Dayhoff

When two proteins descendent from a common ancestor have accumulated a large number of amino acid substitutions in their sequences, it is difficult to establish their common origin. By a careful selection of methods, it is possible to detect very distant relationships by comparisons that rely on statistical evaluations of the similarity between the sequences. These methods all depend on the comparison of an amino acid in one sequence with a corresponding amino acid in another sequence. Each amino acid pair is assigned a numerical value. These are accumulated over a string of residues to give a score. In this chapter we evaluate and compare several scoring systems. The simplest system assigns a value of +1 to identical residues and 0 to nonidentical ones; the scoring matrix corresponding to this system is called the unitary matrix (UM). A slightly more complicated scoring system reflecting the minimum number of base changes required to alter the codon for one amino acid to that for another assigns 3 for amino acid identities, 2 for amino acids whose codons differ by a single base, 1 for amino acids whose codons differ by two bases, and 0 for amino acids whose codons differ in all three bases; we refer to this as the genetic code matrix (GCM). In 1971, a matrix based on alternative amino acids (AAAM) at each position in alignments of groups of related sequences was derived by McLachlan.[1,2] In 1967, we derived a scoring matrix, called the mutation data matrix (MDM_{67}), from 423 accepted point mutations observed in closely related sequences available then.[3] In 1969, we recalculated the mutation data matrix (MDM_{69}) on the basis of 814 accepted point mutations.[4] These matrices were derived in essentially the same manner described in chapter 22.

In our experience, scoring systems representing the average way in which amino acids change during evolution have proved most satisfactory for detecting distant relationships between protein sequences. On the basis of data available through Supplement 2, including 1,572 mutations, we have again recalculated the mutation probability matrices, the odds matrices, and the log odds matrices for various evolutionary distances (see chapter 22). Presumably, in detecting relationships, the best results would be obtained with a matrix corresponding to the same evolutionary distance as that between the sequences being compared. Because we are most interested in obtaining a significant score for comparisons between very distantly related sequences, we will concentrate on matrices that are calculated at large evolutionary distances. In Figure 85, we show the log odds matrix for 250 PAMs to two significant figures; this evolutionary distance corresponds to sequences that are about 80% different. In order to establish its superiority, we compare this 250-PAM matrix with the new mutation data matrices at other evolutionary distances, with other scoring matrices, and with partial information from the matrix itself, using two different computer methods. This extends work reported previously.[5,6]

Measurements of Similarity between Sequences

We currently use two statistical computer methods for assessing the extent of similarity between sequences, programs ALIGN and RELATE, described in chapter 1. ALIGN determines the maximum score that can be achieved by an alignment of a pair of sequences and compares that score with the scores achieved by random permutations of the two sequences. The alignment score is the difference between the score for the real sequences and the average score from the randomized sequences divided by the standard deviation of the scores from the randomized sequences. In order to investigate the dependence of alignment scores on evolutionary distance, we have constructed a model sequence of 100 residues having an average amino acid composition. From this initial sequence, a family of other sequences with known

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A Ala	18																			
R Arg	-15	61																		
N Asn	2	0	20																	
D Asp	3	-13	21	39																
C Cys	-20	-36	-36	-51	119															
Q Gln	-4	13	8	16	-54	40														
E Glu	3	-11	14	34	-53	25	38													
G Gly	13	-26	3	6	-34	-12	2	68												
H His	-14	16	16	7	-34	29	7	-21	65											
I Ile	-5	-20	-18	-24	-23	-20	-20	-26	-24	45										
L Leu	-19	-30	-29	-40	-60	-18	-34	-41	-21	24	59									
K Lys	-12	34	10	1	-54	7	-1	-17	0	-19	-29	47								
M Met	-11	-4	-17	-26	-52	-10	-21	-28	-21	22	37	4	64							
F Phe	-35	-45	-35	-56	-43	-47	-54	-48	-18	10	18	-53	2	91						
P Pro	11	-2	-5	-10	-28	2	-6	-5	-2	-20	-25	-11	-21	-46	59					
S Ser	11	-3	7	1	0	-5	0	11	-8	-14	-28	-2	-16	-32	9	16				
T Thr	12	-9	4	-1	-22	-8	-4	0	-13	1	-17	0	-6	-31	3	13	26			
W Trp	-58	22	-42	-68	-78	-48	-70	-70	-28	-51	-18	-35	-42	4	-56	-25	-52	173		
Y Tyr	-35	-42	-21	-43	3	-40	-43	-52	-1	-9	-9	-44	-24	70	-49	-28	-27	-2	101	
V Val	2	-25	-17	-21	-19	-19	-18	-14	-22	37	19	-24	18	-12	-12	-10	3	-62	-25	43

Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val

Figure 85. Scoring matrix for the evolutionary distance of 250 PAMs. This is the log of the odds matrix; elements are shown multiplied by 100. We refer to this mutation data scoring matrix as MDM_{78}. We have detected no statistically meaningful difference in the results using this matrix and those using the matrix in Figure 84, which has one significant figure less.

numbers of point mutations was generated by the process described in chapter 22. Figure 86 shows the results of these simulation experiments.[7] If we take an alignment score of 3.0 SD as an indication of probable relatedness, these simulations suggest that relatedness to the initial sequence can be demonstrated for sequences that have accumulated 550 mutations in 100 residues and are nearly 88% different. In real comparisons there are additional problems, such as differences in length due to insertions and deletions of genetic material and the non-average behavior of amino acids in any particular molecule. Nevertheless, we can usually detect relationships between real sequences of 100 residues that are 85% different.

The other statistical computer program, RELATE, makes an exhaustive comparison of all segments of a given length from one sequence with those from the other. The average value of a preassigned number of the highest scores is determined. This average value is compared with the distribution of such average values from

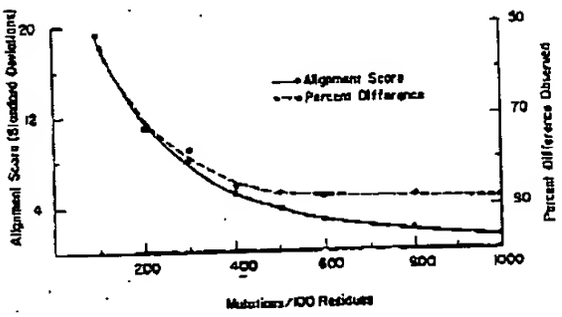

Figure 86. Dependence of alignment scores on the total number of mutations. Pairs of model protein sequences of 100 residues with average amino acid composition were used. A score above 3.0 SD, obtained for pairs of sequences reflecting up to 550 mutations per 100 residues, is considered good evidence of relatedness. The percent difference between sequences, although a good measure at short distances, deteriorates rapidly and approaches an asymptote. This figure is adapted from Ref. 7.

randomized sequences. The segment comparison score is the difference between the average values for the real pair and the mean of the average values from the randomized sequences divided by the standard deviation of the values from the randomized sequences. This program is specifically designed to detect homology between sequences that are very different in length or have only certain portions of their sequences conserved. RELATE can also be used to detect internal duplications in sequences. The alterations in this algorithm that are necessary for detecting internal duplications in a protein are obvious: instead of comparing a sequence with a second sequence, it is compared with itself. Exactly corresponding segments are excluded from the analysis.

Effect of Evolutionary Distance on the Mutation Data Matrices

In order to examine the effect of the mutation distance at which the scoring matrix is derived, we chose six typical pairs of related sequences, ranging between 73% and 86% different from one another. Each pair was tested with 11 matrices. Figure 87 shows how the alignment scores for six pairs of real sequences vary as a function of the evolutionary distance for which the mutation data scoring matrix is computed. The optimum choice of scoring matrix is sequence-dependent; it is a function of factors such as the types of residues that are conserved and the actual evolutionary distance between the proteins. A matrix calculated at a distance of 250 PAMs is close to the optimal choice over this range of protein comparisons. At 250 PAMs, two of the comparison curves are increasing in value, one is decreasing, and three are nearly level. All of the curves are near their maxima and have values greater than 3.5 SD.

Figure 88 shows how the segment comparison scores vary as a function of the evolutionary distance at which the scoring matrix is computed. Four pairs of sequences were identical with those tested using alignment scores. For these, scores are almost 2 SD lower, confirming that the alignment scores are more sensitive for sequences of similar length and architecture. The matrix calculated at 250 PAMs is the optimal choice here; this is consistant with the result obtained with ALIGN, despite the choices of parameters, such as penalty, bias, and segment comparison length, and despite the very different methodologies employed in these two programs. We refer to the 250-PAM mutation data scoring matrix as MDM_{78}.

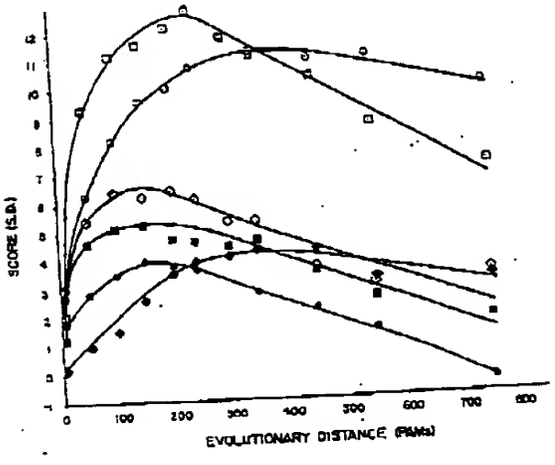

Figure 87. Alignment scores as a function of the evolutionary distance of the mutation data matrices. These log odds matrices, multiplied by 10, were calculated at 4, 50, 100, 150, 200, 242, 300, 350, 450, 550, and 750 PAMs. The gap penalty factor and matrix bias were both given values of 6 in all trials. All scores are based on 300 randomized sequence comparisons, and the standard deviations of the scores are therefore about 4% of their values. The following sequence comparisons were made: *open circle*, hemoglobin alpha chain—human vs. myoglobin—human; *solid circle*, hemoglobin alpha chain—human vs. globin CTT-III—midge larva; *open diamond*, cytochrome c—horse vs. cytochrome c_6—*Spirulina maxima*; *solid diamond*, cytochrome c—horse vs. cytochrome c_{551}—*Desulfovibrio gigas*; *open square*, Ig mu chain C4 homology region—human Gal vs. Ig epsilon chain C4 homology region—human Nd; *solid square*, Ig mu chain C4 homology region—human Gal vs. beta$_2$-microglobulin—human.

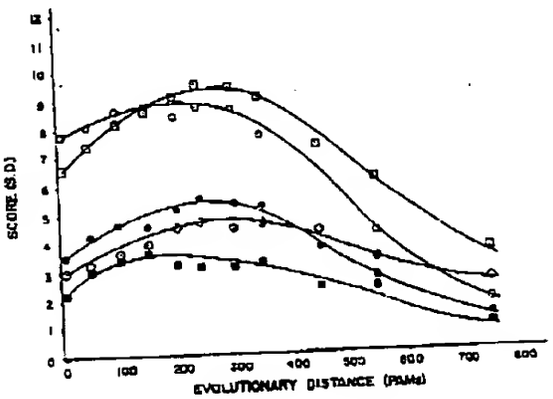

Figure 88. Segment comparison scores as a function of the evolutionary distance of the mutation data matrix. All comparisons were made with a segment length of 20 residues, and all scores are based on 300 randomized sequence comparisons. Matrices were calculated at 4, 50, 100, 150, 200, 242, 300, 350, 450, 550, and 750 PAMs. The sequences compared are the same as for Figure 87 except that the comparison of human myoglobin and midge larva globin is substituted for the comparison of human hemoglobin alpha chain and midge larva globin. Both the human hemoglobin alpha chain–midge larva globin and horse cytochrome c–bacterial cytochrome c_{553} sequence pairs were too distantly related for their scores to be above 3 SD with any matrix.

Comparison of Scoring Matrices

We have compared a number of scoring matrices using ALIGN. The results of these comparisons, involving a broad selection of pairs of related sequences, are listed in Table 24. MDM_{78} gives the highest average score, although it does not always give the highest score for a particular comparison. It is the only matrix that consistently detects relatedness (scores ≥ 3.0 SD) for the entire range of sequences tested. In the comparison of antibacterial substance A with neocarzinostatin, GCM and UM give better scores than MDM_{78}. This may be due to the conservation of what are usually more mutable amino acids. In the other comparisons, matrices based on mutation data perform better, and MDM_{78} usually gives the strongest indication of relatedness. The average scores are shown at the bottom of the table. The score using MDM_{78} is 1 SD better than that using AAAM, 2 SD better than that using GCM and almost 3 SD better than that using UM.

Table 25 shows segment comparison scores for a broad range of sequences including tests for internal duplications using different scoring matrices. Again, on occasion another matrix gives a better score, but only MDM_{78} consistently indicates known relationships between sequences; of the scoring matrices that we tested, it is clearly the best. The average score using it is 2.5 SD better than that for any of the other matrices.

In order to ascertain whether either ALIGN or RELATE produces false-positive results with any of the scoring matrices we tested, we examined 28 pairs of unrelated proteins. Neither program gave false-positive results with any of the matrices. The mean alignment score for the 28 comparisons was between 0.2 and -0.2 for all four matrices. The mean segment comparison score for the 28 pairs was between 0.3 and -0.4 for all four matrices. All of these trials were based on 100 randomized sequence comparisons.

Comparison of MDM_{78} with Its Predecessors

Using a variety of distantly related sequences, we have compared the results using the recently derived MDM_{78}, the two previous mutation data matrices, MDM_{67} and MDM_{68}, based on one-fourth and one-half as much data, respectively, and components of MDM_{78}: the diagonal elements alone, with all off-diagonal elements equal to zero, and the off-diagonal elements, with the diagonal

Table 24
Comparison of Matrices for Calculating Alignment Scores

Sequences Compared	Score (in SD units) Obtained with UM	GCM	AAAM	MDM_{78}
Antibacterial substance A — *Streptomyces* vs. Neocarzinostatin — *Streptomyces*	3.1	3.2	2.6	2.9
Ferredoxin — *Clostridium pasteurianum* vs. Ferredoxin — *Spirulina maxima*	0.1	1.6	1.8	3.4
Hemoglobin alpha — Human vs. Myoglobin — Human	5.8	6.6	9.9	10.7
Hemoglobin alpha — Human vs. Globin CTT-III — Midge larva	2.0	2.4	3.2	3.5
Cytochrome c — Horse vs. Cytochrome c_6 — *Spirulina*	4.5	4.3	7.3	6.1
Cytochrome c — Horse vs. Cytochrome c_{553} — *Desulfovibrio*	0.2	0.4	0.4	3.9
$Beta_2$-microglobulin — Human vs. Ig mu chain C4 homology region — Human Gal	3.6	3.3	4.7	4.8
Ig mu chain C4 homology region — Human Gal vs. Ig epsilon chain C4 homology region — Human Nd	4.7	9.0	9.2	12.1
Average score	3.0	3.9	4.9	5.9

In these comparisons, we used values for the gap penalty (P) and the matrix bias (B) that have been useful for a broad selection of sequence comparisons in our experience, typically 60 and 60 for MDM_{78} (Figure 85), 1 and 1 for GCM, and 0.3 and 0.3 for UM. In the comparison of antibacterial substance A with neocarzinostatin, a bias of 20 and a penalty of 80 were used with MDM_{78} because these are more typical choices for detecting very distant sequence relationships. In the comparisons using AAAM, for which our experience is limited, we varied B from −2 to +4; P was chosen to be 6 and 8. These values produced alignments that were similar in numbers of gaps and gap length to alignments using the other scoring matrices; they produced scores that were statistically indistinguishable from one another. For the above values, we used P = 6 and B = −2. Three hundred randomized sequence comparisons were used in determining scores for AAAM and MDM_{78}; thus, the estimated percent standard deviations of these scores are 4%. UM and GCM scores were calculated using 100 randomized sequence comparisons; thus, the estimated percent standard deviations of these scores are 7%.

Table 25

Comparison of Matrices for Calculating Segment Comparison Scores

Sequences Compared	Score (in SD units) Obtained with UM	GCM	AAAM	MDM_{78}
Cytochrome c_6 — *Monochrysis* vs. Cytochrome c_1 — *Rhodospirillum*	4.7	3.1	2.5	3.5
Azurin — *Bordetella* vs. Plastocyanin — French bean	1.6	2.8	3.1	4.1
Ferredoxin — *Clostridium pasteurianum* vs. Ferredoxin — *Desulfovibrio*	3.9	3.1	4.3	6.0
Troponin C — Rabbit vs. Parvalbumin — Pike	7.6	8.3	8.0	10.2
Troponin C — Rabbit vs. Myosin A1 light chain — Rabbit	8.0	9.3	6.7	15.1
Internal Duplication				
Tropomyosin alpha chain — Rabbit	5.9	4.0	3.6	8.3
Protease inhibitor, submandibular gland — Dog	4.1	3.6	5.3	7.9
Cytochrome c_3 — *Desulfovibrio gigas*	0.5	1.3	0.7	3.9
Ferredoxin — *C. pasteurianum*	7.8	5.9	7.1	7.7
Average score	4.9	4.6	4.6	7.4

In the cytochrome c_2-c_3 and in the ferredoxin internal duplication comparisons, a segment length of 15 residues was used; in the other comparisons, we used a segment length of 20 residues. Three hundred randomized sequence comparisons were used in calculating scores for AAAM and MDM_{78}; thus, the percent standard deviations for these scores are 4%. One hundred comparisons were used for UM and GCM; thus, their percent standard deviations are 7%.

Table 26

Comparison of Mutation Data Matrices for Calculating Alignment Scores

Sequences Compared	Scores (in SD units) Obtained with MDM_{87}	MDM_{69}	MDM_{78}	UM	Diagonal Only MDM_{78}	Off-diagonal and Averaged Diagonal MDM_{78}
Antibacterial substance A — *Streptomyces* vs. Neocarzinostatin — *Streptomyces*	2.0	2.4	2.9	3.1	1.4	1.8
Ferredoxin — *Clostridium pasteurianum* vs. Ferredoxin — *Spirulina maxima*	2.6	2.6	3.4	0.1	2.7	2.7
Hemoglobin alpha — Human vs. Myoglobin — Human	9.9	9.7	10.7	5.8	9.9	10.3
Hemoglobin alpha — Human vs. Globin CTT-III — Midge larva	2.6	2.4	3.5	2.0	0.9	3.5
Cytochrome c — Horse vs. Cytochrome c_6 — *Spirulina*	5.6	5.4	6.1	4.5	5.6	5.8
Cytochrome c — Horse vs. Cytochrome c_{553} — *Desulfovibrio*	3.8	3.9	3.9	0.2	2.0	2.8
$Beta_2$-microglobulin — Human vs. Ig mu chain C4 homology region — Human Gal	3.3	2.8	4.8	3.6	3.9	4.8
Ig mu chain C4 homology region — Human Gal vs. Ig epsilon chain C4 homology region — Human Nd	10.1	11.5	12.1	4.7	11.2	11.9
Average score	5.0	5.1	5.9	3.0	4.7	5.5

In the comparison of antibacterial substance A with neocarzinostatin, a matrix bias of 20 and a gap penalty of 80 were used with MDM_{78} and its derivatives. In the other comparisons with MDM_{78}, a bias of 60 and a penalty of 60 were used. A penalty of 8 and a bias of 6 were used with MDM_{69} and MDM_{87} because their elements are expressed to one significant figure less than MDM_{78}. Three hundred random comparisons were used in determining scores for all matrices except UM, for which 100 random comparisons were made. A penalty of 0.3 and bias of 0.3 were used with UM.

Table 27

Comparison of Mutation Data Matrices for Calculating Segment Comparison Scores

Sequences Compared	Scores (in SD units) Obtained with MDM_{67}	MDM_{69}	MDM_{78}	UM	Diagonal Only MDM_{78}	Off-diagonal and Averaged Diagonal MDM_{78}
Cytochrome c_6 — *Monochrysis* vs. Cytochrome c_2 — *Rhodospirillum*	2.9	2.6	3.5	4.7	3.0	3.2
Azurin — *Bordetella* vs. Plastocyanin — French bean	3.8	3.7	4.1	1.6	2.5	3.2
Ferredoxin — *Clostridium pasteurianum* vs. Ferredoxin — *Desulfovibrio*	5.4	5.3	6.0	3.9	5.2	3.9
Troponin C — Rabbit vs. Parvalbumin — Pike	10.0	9.8	10.2	7.6	4.8	11.8
Troponin C — Rabbit vs. Myosin A1 light chain — Rabbit	14.8	13.4	15.1	8.0	9.0	13.3
Internal Duplication						
Tropomyosin alpha chain — Rabbit	7.8	6.9	8.3	5.9	4.7	8.8
Protease inhibitor, submandibular gland — Dog	6.5	5.4	7.9	4.1	4.2	6.6
Cytochrome c_3 — *Desulfovibrio gigas*	1.7	3.8	3.9	0.5	3.2	1.6
Ferredoxin — *C. pasteurianum*	7.1	7.3	7.7	7.8	7.3	7.5
Average score	6.7	6.5	7.4	4.9	4.9	6.8

We used a segment length of 16 residues for the cytochrome c_2-c_6 and ferredoxin internal duplication comparisons and 20 residues for the other comparisons. Three hundred randomized sequence comparisons were used in determining scores for matrices except UM, for which 100 randomized sequence comparisons were used.

elements equal to 60 (the approximate average value for diagonal elements in MDM_{78}). These components can be thought of as intermediate between UM and MDM_{78}. The first has zero for all off-diagonal elements; however, the pattern of MDM_{78} is retained in the diagonal elements. In the other modification, all of the diagonal elements are equal as in UM, but the nondiagonal elements from MDM_{78} are retained. In Table 26, the results from program ALIGN are shown, and in Table 27, the results from program RELATE are shown. MDM_{78} is clearly superior to the earlier ones, which were based on less data. The main differences are in the off-diagonal elements, particularly the negative ones for which there was previously little data.

Both amino acid mutabilities (diagonal elements) and their exchange probabilities (off-diagonal elements) are important aspects of MDM_{78}. Comparison of the results using the diagonal elements only with those using UM shows that the pattern in the diagonal elements is helpful in calculating alignment scores. In calculating segment comparison scores, the diagonal elements are not, on the average, an improvement over UM. However, diagonal elements are helpful in some cases, such as in detecting the internal duplication in cytochrome c_3. Comparison of UM with the matrix containing the off-diagonal and the averaged diagonal elements shows that the off-diagonal terms contribute more to the good results than the variability in the diagonal terms. The MDM_{78} that contains both patterns is clearly superior to either of the partial matrices.

References

1. McLachlan, A.D., J. Mol. Biol. 61, 409–424, 1971
2. McLachlan, A.D., J. Mol. Biol. 64, 417–437, 1972
3. Dayhoff, M.O., and Eck, R.V., Atlas of Protein Sequence and Structure, Vol.3, pp.33–41, Nat. Biomed. Res. Found., Silver Spring, Md., 1968
4. Dayhoff, M.O., Eck, R.V., and Park, C.M., *in* Atlas of Protein Sequence and Structure, Vol.5, ed. Dayhoff, M.O., pp.89–99, Nat. Biomed. Res. Found., Washington D.C., 1972
5. Schwartz, R.M., and Dayhoff, M.O., *in* Evolution of Protein Molecules, ed. Matsubara, H., and Yomenaka, T., pp.1–16, Japan Sci. Soc. Press, Tokyo, 1978
6. Schwartz, R.M., and Dayhoff, M.O., *in* Origin of Life: Proceedings of the Second ISSOL Meeting, the Fifth ICOL Meeting, ed. Noda, H., pp.457–469, Japan Sci. Soc. Press, Tokyo, 1978
7. Dayhoff, M.O., Barker, W.C., and McLaughlin, P.J., Origins Life 5, 311–330, 1974

24 Duplications in Protein Sequences

W.C. Barker, L.K. Ketcham, and M.O. Dayhoff

We have reported a routine procedure for screening protein sequences for evidence of intragenic duplications;[1] at that time we tested 163 protein sequences representing all of the 116 superfamilies of unrelated proteins listed in the *Atlas*, Supplement 2, chapter 2. Twenty superfamilies were found to contain proteins that reflect internal gene duplications. The intragenic duplications detected were of two major types: (1) one or more duplications of all or part of a gene to produce a protein with two or several regions of detectable sequence homology, and (2) repeated reduplication of a small DNA segment to produce a protein that is repetitive over most of its length. These duplications had occurred in prokaryotes and eukaryotes over a wide span of evolutionary history, from perhaps 3 billion years ago in an early anaerobic bacterium (clostridial-type ferredoxin) to very recently in the human line since the divergence of other primates (the haptoglobin alpha-2 chain). We have extended this study by testing 117 of the sequences that appear in this volume, one from each of the new protein superfamilies and families listed in Table 2.

Computer Method

The computer program RELATE, described in chapter 1, can be used to detect repeated patterns within a protein sequence. The program compares every possible segment of a given length with every other segment of that length within the sequence. A segment score is accumulated from comparisons of the amino acids occupying corresponding positions within the two segments. No gaps are permitted within the segments being compared. A matrix of comparison scores for each amino acid pair is supplied.

A numerical property of the distribution of segment scores is determined for the real sequence and for at least 100 permuted sequences with the same amino acid composition. The segment comparison score is calculated as the difference between the value determined for the real sequence and the average value determined from all of the permuted sequences, divided by the standard deviation of the values from the permuted sequences. The segment comparison score is thus expressed in SD units, and the probability of occurrence by chance of a score higher than a particular value can be obtained from the cumulative standardized normal distribution table. A score ≥ 3.0 SD ($P < 0.0014$) is taken as indicative of internal duplication. For the numerical property, we have used the mean of a predetermined number of highest segment scores. This was determined for each protein from the segment length (s) and the total length of the sequence (L): $L/2 - s + 1$. This expression is equal to the number of scores to be expected from comparisons of corresponding segments if the sequence has exactly doubled.

Several other kinds of output are obtained from the computer program. An ordered list of many segment comparisons giving the highest scores is printed, from which the regions of the sequence showing unusual similarity can be identified. From the list, a table of displacements of the matching segments, giving the frequency of occurrence and average score for each displacement, is constructed. A protein that has duplicated will have many high scores at a displacement of half of its total length. A protein with a prominent 10-residue periodicity will have many high scores at displacements of 10, 20, 30, etc.

Although the segment comparison score provides an easy and straightforward criterion of internal duplication, it may fail to detect duplications in certain cases. If a duplication involves only a small fraction of the sequence, it may not be detected unless it is composed of amino acids that are usually highly conserved in proteins. We have purposely designed the procedure to detect major duplication and extensive periodicity. If too many changes have occurred in the sequence, an ancestral duplication may not be detected.

Table 28
Proteins with Previously Detected Duplications

	Score[a] (SD units)	Length of Protein	Approximate Length of Repeat	Number of Repetitions	Percent Repetitious
Eukaryote Sequences					
Collagen alpha 1 chain (rat)	13.3[b]	1,052	3	337	96%
Lipid-binding protein A-I (human)	11.4	245	11	18	81%
Keratin B2A (sheep)	8.6	171	10	13	76%
Keratin B-IIIA3 (sheep)	4.0	131	10	11	84%
Alpha$_{S1}$ casein (sheep)	3.3[c]	199	20	>4	?
Bombinin (unk)	3.8[d]	24	4	4	67%
Immunoglobulin mu chain C region (human)	19.1	452	108	4	96%
Immunoglobulin epsilon chain C region (human)	15.3	423	108	4	100%
Immunoglobulin gamma chain C region (guinea pig)	9.4	329	108	3	98%
Serum albumin (human)	12.1[b]	584	195[e]	3	100%
Sperm histone (bovine)	3.7[c]	47	8	3	51%
Histone H3 (bovine)	3.6	135	{9 13	{3 2	39%
Haptoglobin alpha 2 chain (human)	37.1	143	59	2	83%
Troponin C, skeletal muscle (rabbit)	7.5	159	76[e]	2	96%
Myosin A1 light chain (rabbit)	7.6[f]	190	76[e]	2	80%
Parvalbumin (pike)	3.6	108	39	2	72%
Lipid-binding protein C-I (human)	3.2[g]	57	27[e]	2	95%
Prothrombin (bovine)	10.8[h]	582	79	2	27%
Neurophysin 2 (pig)	4.9	92	23	2	50%
Posterior pituitary peptide (bovine)	4.0[i]	48	11	2	46%
Alpha crystallin A chain (bovine)	4.0[i]	173	30	2	35%
Protease inhibitor, Bowman-Birk (soybean)	3.9	71	28	2	79%
Prokaryote Sequences					
Cytochrome c_3 (*Desulfovibrio desulfuricans*)	4.3[d]	102	17	4	67%
Cytochrome c_7 (*Desulfuromonas acetoxidans*)	4.1	68	18	3	79%
Cytochrome c_3 (*Desulfovibrio vulgaris*)	3.7	107	50[e]	2	93%
Murein-lipoprotein (*Escherichia coli*)	8.2	58	14[e]	2.5	60%
Ferredoxin (*Clostridium pasteurianum*)	6.8	55	28	2	100%
Ferredoxin (*Chromatium* sp.)	3.3[j]	81	28	2	69%
Rubredoxin (*Pseudomonas oleovorans*)	3.5	174	55	2	63%

Modified from Table 5 in Barker, W.C., Ketcham, L.K., and Dayhoff, M.O., J. Mol. Biol. 10, 265-281, 1978.

[a]Unless otherwise noted, all scores were determined using the 1978 mutation data matrix, a segment length of 20, and 100 random runs.

[b]Score based on testing residues 1-500.

[c]Score obtained using segment length of 16 and 300 random runs.

[d]Score obtained using segment length of 8.

[e]This major repeating unit shows evidence of repetitious structure within itself.

[f]Score based on testing residues 31-190.

[g]Score obtained using segment length of 4 and 300 random runs.

[h]Score based on testing residues 1-323.

[i]Score obtained using segment length of 12.

[j]Score obtained using segment length of 8 or 12 and 300 random runs.

In this study, the segment length used initially was 15 for sequences 50 residues or longer, 12 for sequences of 30-49 residues, and 8 for those less than 30. A number of the sequences were further tested using longer or shorter segment lengths. Sequences that are long and have had no deletions or insertions since duplication give higher segment comparison scores with longer segment lengths. If many insertions or deletions have occurred since the duplication, reducing the segment length may increase the segment comparison score. The scoring matrix used was the mutation data matrix shown in Figure 84.

Results

In Table 28 we have listed the proteins found previously[1] to have duplications detectable by this method. The nine additional proteins with duplications detected in the present investigation are shown in Table 29. Four of these represent new families in superfamilies containing previously known duplicated proteins. The five kringle regions in the amino-terminal portion of plasminogen are very similar to the two kringle regions of prothrombin (see Alignment 9 and Figure 14). Calcium-dependent regulator protein and myosin DTNB light chain share the two successive internal duplications found also in the structures of troponin C and myosin alkali light chains (see Alignment 35 and Figure 57). The immunoglobulin alpha chain C region reflects a history of internal duplications at least partially shared with other immunoglobulin heavy chains (see Alignment 30 and Figures 45 and 46).

Three of the proteins belong to superfamilies that previously did not contain examples of detectable duplications. Submandibular gland protease inhibitor and ovomucoid share the duplication that produced a double-length sequence compared with the related pancreatic secretory trypsin inhibitor, which contains only one homology region. Subsequently, a partial duplication added a third homology region to the ovomucoid sequence (see Alignment 19 and Figure 35). The high potential iron-sulfur protein from *Rhodopseudomonas gelatinosa* is distantly related over its entire length to proteins from *Chromatium vinosum* and *Thiocapsa pfennigii*; nevertheless, duplications were not detected in these sequences. In the *R. gelatinosa* sequence, residues 33-46, 47-60, and 61-74 can be aligned, without gaps, so that three residues are common to all three segments. Altogether, the first two segments share six identities in fourteen residues and each of these segments shares four identities with the third segment. When the other two high potential iron-sulfur protein sequences are aligned with that from *R. gelatinosa*, they are seen to have several insertions that disrupt the repetitive pattern.

The remaining two proteins, tropomyosin and troponin T, represent new superfamilies. The prominent repeating units of tropomyosin are 42 residues long, but they contain a pattern of six 7-residue repeats. Thus, tropomyosin joins collagen, lipid-binding protein A-I, and the two keratins in having a large number of repeats of a short segment. Troponin T may at one time have had a repetitive sequence of this type, but at present no simple pattern of

Table 29

New Proteins with Detected Duplications

Sequence	Score[a] (SD units)	Length of Protein	Approximate Length of Repeat	Number of Repetitions	Percent Repetitious
Plasminogen (human)	44.5	790	79	5	50%
Ovomucoid (Japanese quail)	17.4	186	59	3	95%
Protease inhibitor, submandibular gland (dog)	7.4	115	54	2	94%
Calcium-dependent regulator protein (bovine)	13.0	148	74[b]	2	100%
Myosin DTNB light chain (rabbit)	3.3[c]	169	76[b]	2	90%
Tropomyosin alpha chain (rabbit)	5.8	284	42[b]	7	100%
Troponin T, skeletal muscle (rabbit)	3.5[c]	259	?	?	?
Immunoglobulin alpha-1 chain C region (human)	8.7	353	108	3	92%
High potential iron-sulfur protein (*Rhodopseudomonas gelatinosa*)	6.1	74	14	3	57%

[a]Unless otherwise noted, scores were determined with a segment length of 15 and 100 random runs.

[b]This major repeating unit shows evidence of repetitious structure within itself.

[c]Score obtained using segment length of 25.

Table 30
Superfamilies Containing Duplicated Proteins

Superfamily No.	Superfamily Name	Families Total	Families Detectable Duplications	Families Not Duplicated	Families Probable Duplications
2	Cytochrome c_3 related	3	3		
6	Ferredoxin related	9	2		7
7	High potential iron-sulfur protein	3	1		2
8	Rubredoxin	2	1	1	
39	Prothrombin related	9	2	7	
54	Protease inhibitors (PSTI-type)	5	2	3	
58	Protease inhibitors (Bowman-Birk type)	3	1	2	
64	Posterior pituitary peptide	1	1		
81	Hemolytic peptides	2	1		1
89	Immunoglobulin C regions and related proteins	7	4	3	
94	Histone H3	1	1		
97	Sperm histone	1	1		
139	Alpha crystallin	1	1		
141	Keratin high-sulfur fraction B2 related	2	2		
145	Collagen	1	1		
147	Tropomyosin	1	1		
149	Troponin C related	6	5		1
150	Troponin T	1	1		
152	Animal lipid-binding proteins	4	2		2
154	Murein-lipoprotein	1	1		
155	Alpha$_{S1}$ casein	1	1		
161	Neurophysins	1	1		
164	Haptoglobin alpha chain	1	1		
169	Serum albumin	1	1		

duplications can explain the positions of the high-scoring matches. However, the sequence contains an unusual number of acidic and basic residues (see Table 33) consistent with a history of gene duplication.

In Table 30 we have listed all of the superfamilies that contain proteins with detected duplications. In most cases where more than one family (a group of proteins that are less than 50% different in sequence) is represented in a superfamily, the duplication is either not detectable or not present in some of the families. Sequences from more than one family are known in 12 of the superfamilies containing duplicated proteins; altogether 56 families are represented in these 12 superfamilies. Clearly distinguishable duplicated regions are seen in 26 of these families. In the other 29, 16 of the families lack the duplication, whereas 13 have sequences of approximately the same length as the duplicated sequences and show weak evidence of duplication, but the sequence information has been degraded by accumulated point mutations, insertions, and deletions. In all, 38 (about 12%) of the 314 families contain proteins with detected duplications.

References

1. Barker, W.C., Ketcham, L.K., and Dayhoff, M.O., J. Mol. Biol. 10, 265-281, 1978.

25 Composition of Proteins

M.O. Dayhoff, L.T. Hunt, and S. Hurst-Calderone

Because it is of some interest to compare the amino acid compositions of the various types of proteins in this *Atlas* and to classify new sequences on the basis of amino acid composition and length, we have prepared the summary shown in Table 33 from which Tables 31, 32, and 34 are obtained. Table 33 is derived from the well-characterized sequences named in the Superfamily List of chapter 2 and described in the *Atlas of Protein Sequence and Structure 1972*, Volume 5 and Supplements 1, 2, and 3. These were grouped into 314 families of sequences less than 50% different from one another. One sequence from each entry (containing at least 20 residues and differing by at least 5% from other entries) is included. The immunoglobulin constant and variable region sequences are treated as separate entries. Seven sequences contributed to the proinsulin values; the undetermined residues at the ends of the C-peptides were assumed to be three arginines and one lysine, as in the human sequence. No complete sequence of the collagen alpha 1 chain from any one organism has been determined; however, a composite sequence of all 1,052 residues was derived by combining the almost completely determined bovine sequence with residues 140–418 of the rat sequence (see chapter 2).

From the total number of residues of each kind for all sequences in the family, the percentages of groups of similar amino acids, such as basic, acidic, aromatic, and large hydrophobic amino acids, were calculated. These are shown for each family of proteins in Table 33. The length shown is that of the first protein of the family on our data tape. The families are ordered approximately as they occur on the data pages so that distantly related groups are adjacent to each other. An alphabetized index of family names is shown following Table 33.

An average composition of the sequenced proteins was determined (see Table 31). The average percent of amino acid i, F_i, is given by the following equation:

$$F_i = \frac{\sum_{j=1}^{314} L_j f_{ij}}{\sum_{j=1}^{314} L_j}$$

where

L_j is the length of a typical protein in family j,
f_{ij} is the % of amino acid i in family j, and
j is summed over the 314 families.

This average is the percent composition for the residues in a pool of 314 sequences, each with typical length and average percent composition for its family. The contribution of certain groups of residues to the percent compositions is shown in Table 32.

Table 31
Average Percent of Amino Acids in Proteins

Ala	A	8.6	Arg	R	4.9
Gly	G	8.4	Ile	I	4.5
Leu	L	7.4	Asn	N	4.3
Ser	S	7.0	Gln	Q	3.9
Val	V	6.6	Phe	F	3.6
Lys	K	6.6	Tyr	Y	3.4
Thr	T	6.1	Cys	C	2.9
Glu	E	6.0	His	H	2.0
Asp	D	5.5	Met	M	1.7
Pro	P	5.2	Trp	W	1.3

The percent composition of a pool of 314 sequences, one for each family in Table 33, is shown in decreasing order of occurrence of the amino acid. Undetermined amides occurred in 0.5% of the positions. These were allocated equally to Asp, Asn, Glu, and Gln.

Table 32
Average Percent of Amino Acid Groups in Proteins

Small aliphatic	A+G	16.9
Hydroxyl	S+T	13.1
Acidic	D+E	11.6
Acidic + acid amide	D+B+N+E+Z+Q	19.8
Basic	K+R+H	13.5
Hydrophobic	L+V+I+M	20.2
Aromatic	F+Y+W	8.3

Table of Contents

LIST OF FIGURES